Nondecomposable Data Dependent Regularizers offer Significant Performance Gains

Anonymous authors

I. Introduction

Data dependent regularization is known to benefit a wide variety of problems in machine learning. Often, these regularizers cannot be easily decomposed into a sum over a finite number of terms, e.g., a sum over individual example-wise terms. The F_β measure, Area under the ROC curve (AUCROC) and Precision at a fixed recall (P@R) are some prominent examples that are used in many applications. We find that for most medium to large sized datasets, scalability issues severely limit our ability in leveraging the benefits of such regularizers. Importantly, the key technical impediment despite some recent progress is that, such objectives remain difficult to optimize via backpropapagation procedures. While an efficient general-purpose strategy for this problem still remains elusive, in this paper, we show that for many data-dependent nondecomposable regularizers that are relevant in applications, sizable gains in efficiency are possible with minimal code-level changes; in other words, no specialized tools or numerical schemes are needed. Our procedure involves a reparameterization followed by a partial dualization – this leads to a formulation that has provably cheap projection operators. We present a detailed analysis of runtime and convergence properties of our algorithm. On the experimental side, we show that a direct use of our scheme significantly improves the state of the art IOU measures reported for MSCOCO Stuff segmentation dataset.

A. Why are Nondecomposable regularizers relevant?

Consider the situation where we would like to ensure that the performance of a statistical model is uniformly good over groups induced via certain protected attributes (such as race or gender). Or alternatively, we may want that when updating an algorithm in a manufacturing process, the new system's behavior should mostly remain similar with respect to some global measures such as makespan [1]. And finally, when pooling datasets from multiple sites, global characteristics of Precision-Recall should be (approximately) preserved across sites [2]. Motivated by these issues encountered in various real world problems, recently [3] presents a comprehensive study of the computational aspects of learning problems with *rate constraints* – there, the constrained setting is preferred due to several reasons including its ease of use for a practitioner. The authors show that for a general class of constraints, a proxy-Lagrangian based method must be used because the Lagrangian is not optimal. This raises the question whether there exist a broad class of data-dependent *nondecomposable* functions for which the regularized/penalized formulation based on *standard* Lagrangian schemes may, in fact, be effective and sufficient.

B. Our contributions

We first *reparameterize* a broad class of nondecomposable data-dependent regularizers into a form that can be efficiently optimized using first order methods. Interestingly, this reparameterization naturally leads to a Lagrangian based procedure where existing SGD based methods can be employed with little to no change. While recent results suggest that optimizing nondecomposable data-dependent regularizers may be challenging, our development shows that a sizable subclass of such regularizers indeed admit simple solution schemes. Our overall procedure comes with convergence rate guarantees and optimal per-iteration complexity. On the MSCOCO stuff segmentation dataset, we show that a direct use of this technique yields significant improvements to the state of the art, yielding a mean IoU of 0.32. These results have been submitted to the leaderboard.

1) Our Model: Let us write down a formulation which incorporates a data-dependent or shape regularizer. The objective function is a sum of two terms: **(1)** a *decomposable* Empirical Risk Minimization (ERM) term to learn the optimal function which maps x to y and **(2)** a *nondecomposable, data dependent* $\mathcal{S}-$measure regularizer term. In particular, for a fixed $\alpha > 0$, we seek to solve the following optimization problem,

$$\min_{W} \quad \overbrace{\frac{1}{N}\sum_{i=1}^{N} \text{loss}\left(W; x_i, y_i\right)}^{\text{ERM}} \; + \; \alpha \overbrace{\mathcal{R}\left(W; \varphi \circ \hat{D}\right)}^{\mathcal{S}-\text{Measure}}, \qquad (1)$$

where $\hat{D}$ represents the empirical distribution, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}^k$ with $i = 1, \ldots, N$ denoting training data examples, and $\varphi \circ \hat{D}$ represents the shape random variable. We let $\text{loss}(\cdot)$ to be any standard loss function such as a cross entropy loss, hinge loss and so on. **In this work, we will assume that the $\mathcal{S}-$Measure is given by F_β metric** [4] while noting that the results apply directly to other measures such as R@P, P@R and can be easily generalized to other nondecomposable metrics such as AUCROC, AUCPR, using the Nyström method with no additional computational overhead.

2) Theoretical Result: Informally, we show that there exists a generalized reformulation of the linear program in [5] for training deep networks using backpropagation. We then analyze the convergence of our proposed method called as Reparameterized Dual Ascent to solve the problem in (1). Our result can be summarized as follows:

Theorem 1. *Assume that* $\|\nabla_W loss(W; x_i, y_i)\|_2 \leq G_1$, *and* $Var_i\left(\|\nabla_W loss(W; x_i, y_i)\|_2\right) \leq \sigma$ *in the ERM term in Prob.* (1). *Then, our algorithm converges to a* $\epsilon-$*approximate (local) solution of Prob.* (1) *in* $O(1/\sqrt{T})$ *iterations such that for a batch size of* B, *each iteration requires* only $O(B \log B)$.

3) Empirical Result: We applied our algorithm for the problem of semantic segmentation, see [6], [7]. On the MSCOCO stuff segmentation dataset, we show that a direct use of this technique yields significant improvements to the state of the art, yielding a mean IoU of 0.32. These results have been submitted to the leaderboard.

II. Conclusion

We showed how various nondecomposable regularizers may indeed permit highly efficient optimization schemes that can also directly benefit from the optimization routines implemented in mature software libraries used in vision and machine learning. We provide a technical analysis of the algorithm and show that the procedure yields state of the art performance for a semantic segmentation task, with only minimal changes in the optimization routine.

REFERENCES

[1] G. B. Limentani, M. C. Ringo, F. Ye, M. L. Bergquist, and E. O. McSorley, "Beyond the t-test: statistical equivalence testing," 2005.

[2] H. H. Zhou, Y. Zhang, V. K. Ithapu, S. C. Johnson, V. Singh, *et al.*, "When can multi-site datasets be pooled for regression? hypothesis tests, ℓ_2-consistency and neuroscience applications," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4170–4179, JMLR. org, 2017.

[3] A. Cotter, H. Jiang, S. Wang, T. Narayan, M. Gupta, S. You, and K. Sridharan, "Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals," *arXiv preprint arXiv:1809.04198*, 2018.

[4] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal thresholding of classifiers to maximize f1 measure," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer.

[5] E. E. Eban, M. Schain, A. Mackey, A. Gordon, R. A. Saurous, and G. Elidan, "Scalable learning of non-decomposable objectives," *arXiv preprint arXiv:1608.04802*, 2016.

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.

[7] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review*.